

Multiset Codes for Permutation Channels

Mladen Kovačević and Dejan Vukobratović

Department of Power, Electronics, and Communication Engineering,

University of Novi Sad, 21000 Novi Sad, Serbia

Emails: {kmladen, dejanv}@uns.ac.rs

Abstract—This paper introduces the notion of *multiset codes* as relevant to the problem of reliable information transmission over permutation channels. The motivation for studying permutation channels comes from the effect of out of order delivery of packets in some types of packet networks. The proposed codes are a generalization of the so-called *subset codes*, recently proposed by the authors. Some of the basic properties of multiset codes are established, among which their equivalence to integer codes under the Manhattan metric. The presented coding-theoretic framework follows closely the one proposed by Kötter and Kschischang for the operator channels. The two mathematical models are similar in many respects, and the basic idea is presented in a way which admits a unified view on coding for these types of channels.

I. INTRODUCTION

In this paper, we study the problem of error correction in the *permutation channels*. We aim to present a coding-theoretic framework for such channels, which is based on the notion of *multiset codes*. These codes are a generalization of the so-called *subset codes* recently proposed by the authors [1], and are argued to be appropriate constructs for reliable information transmission over permutation channels.

Permutation channels arise, for example, as models for an end-to-end transmission in some types of packet networks. Namely, certain network protocols provide no guarantees on the in-order delivery of packets [2], and in addition to dropping some packets, delivering erroneous packets, etc., have the effect of delivering an essentially random permutation of the packets sent. Examples include a number of recently popular networking technologies such as mobile ad-hoc networks, vehicular networks, delay tolerant networks, wireless sensor networks, etc. In the following section we will give a more detailed description of the channel model that we consider, as well as the basic idea underlying the definition of codes for such a channel. This idea is the same as the one presented by Kötter and Kschischang in their seminal paper [3], which gave rise to the definition of the operator channel as an appropriate model of random linear network coded networks, and codes in projective spaces as adequate constructs for such a channel.

In Section III we will give an overview of the subset coding approach presented in [1]. This approach is then extended and generalized by introducing the so-called multiset codes in Section IV. Some basic properties of multiset codes and their advantages over subset codes are also described in this section. Finally, Section V provides two simple, but fairly general examples of both types of codes.

II. THE CHANNEL MODEL

Let S be a finite alphabet with $|S| = q > 0$ symbols. Without loss of generality, we assume that $S = \{1, 2, \dots, q\}$. By a *permutation channel* over S we understand the channel whose inputs are sequences of symbols from S , and which, for any input sequence, outputs a random permutation of this sequence. As noted in the Introduction, such channels arise in some types of packet networks in which the packets comprising a single message are routed separately and are frequently sent over different routes in the network. Therefore, the receiver cannot rely on them being delivered in any particular order.

In addition to random permutations, the channel can have other deleterious effects on the transmitted sequence, such as *insertions*, *deletions*, and *substitutions* of symbols. Substitutions (i.e., errors) are random alterations of symbols, usually caused by noise. Insertions and deletions can be thought of as synchronization errors, where a symbol is read twice, or is skipped, because of the incorrect timing of the receiver's clock. There are also various other situations where they occur (see, e.g., [4]). For example, in a networking scenario mentioned above, packet deletions can be caused by network congestion and consequent buffer overflows in the routers. Note that, as the transmitted sequence is being permuted, *erasures* are essentially the same as deletions, because the position of the erased symbol (in the original sequence) cannot be deduced. To conclude, the channel considered in this paper is the permutation channel with insertions, deletions, and substitutions.

Remark 1: In the case when the permutation channel models a packet network, it should be pointed out that the framework proposed here assumes an end-to-end network transmission model, and consequently, that coding is done on the transport or application layer. It is a frequent assumption in this scenario that only deletions can occur in the channel (apart from permutations). Namely, it is understood that errors are addressed by error-detecting and error-correcting codes at the lower layers (link and physical layer).

A. Coding for the permutation channel

We now state the main idea in a somewhat informal way; the precise definitions are given in subsequent sections.

Codes for various types of channel impairments that we consider (insertions, deletions, and substitutions) have been thoroughly studied in the literature; but how does one deal with random permutations of the symbols? One solution relies on

the following simple idea: Information should be encoded in an object which is invariant under permutations. An example of such an object is a *set*. Based on this observation, the authors have introduced the so-called subset codes as relevant for the above channel model [1], the codewords of which are taken to be subsets of an alphabet S . An appropriate metric is specified on the space of all subsets of S , after which the definition of codes and their parameters follows familiar lines. In the present paper we further generalize this idea by noting that there exists an even more general object invariant under permutations – a *multiset*. Informally, a multiset is a set with repetitions of elements allowed. Clearly, for a given alphabet S , there are more multisets of certain cardinality than there are sets of the same cardinality, and hence, this approach can increase the code rate, among other advantages.

To conclude this section, we note that we have adopted the above principle of taking codewords to be objects invariant under the channel transformation, from the work of Kötter and Kschischang [3]. These authors have noticed that this principle can be applied to the channels arising in networks which are based on random linear network coding (RLNC). In such networks, random linear combinations of the injected packets are delivered to the receiver, and hence, the only property preserved by such a channel is the vector space spanned by those packets¹. From this observation, the authors of [3] have developed the notion of *subspace codes*, i.e., codes in projective spaces, where codewords are taken to be vector subspaces of some ambient vector space (the space of all packets). There are many parallels between subspace and subset/multiset codes, as will be evident from the exposition in the subsequent sections. In fact, multiset codes can be thought of as a generalization of subspace codes in the sense that any basis of a subspace of the set of all packets is also a subset of the set of all packets. This is a consequence of the fact that the RLNC channel is more restrictive than the permutation channel. Namely, permuting the packets is a special case of delivering multiple linear combinations of the packets.

III. SUBSET CODES

Let S be a nonempty finite set representing the alphabet of the given permutation channel. If this channel models a packet network, we can think of S as the set of all possible packets. Let $\mathcal{P}(S)$ denote the power set of S , i.e., the set of all subsets of S , and $\mathcal{P}(S, \ell)$ the set of all subsets of S of cardinality ℓ .

Definition 1: A *subset code* over an alphabet S is a non-empty subset of $\mathcal{P}(S)$. If $\mathcal{C} \subseteq \mathcal{P}(S, \ell)$, we say that \mathcal{C} is a constant-cardinality code.

As usual, in order to enable the receiver to recover from errors, erasures, etc., codewords of the code should be chosen to differ from each other as much as possible. A measure of “dissimilarity” of sets is needed for this purpose. A natural one, which is in fact a metric on $\mathcal{P}(S)$, is given by:

$$d(X, Y) = |X \triangle Y| \quad (1)$$

¹ Actually, it is preserved only if the transformation applied to the packets is full-rank, but this happens with high probability if the linear combinations are indeed random.

for $X, Y \in \mathcal{P}(S)$, where \triangle denotes the symmetric difference between sets which is defined as $X \triangle Y = (X \setminus Y) \cup (Y \setminus X)$. We also have:

$$\begin{aligned} d(X, Y) &= |X \cup Y| - |X \cap Y| \\ &= |X| + |Y| - 2|X \cap Y| \\ &= 2|X \cup Y| - |X| - |Y|. \end{aligned} \quad (2)$$

The distance $d(X, Y)$ is the length of the shortest path between X and Y in the Hasse diagram [5] of the lattice of subsets of S ordered by inclusion. This diagram plays a role similar to the Hamming hypercube for the classical codes in the Hamming metric, and it is in fact isomorphic to the Hamming hypercube, as discussed below. For constant-cardinality codes, the distance between codewords is always even.

The minimum distance of the code can now be defined as:

$$\min_{X, Y \in \mathcal{C}, X \neq Y} d(X, Y). \quad (3)$$

Other important parameters of the code are its size $|\mathcal{C}|$, maximum cardinality of the codewords:

$$\max_{X \in \mathcal{C}} |X|, \quad (4)$$

and the cardinality of the ambient set, $|S|$. The code $\mathcal{C} \subseteq \mathcal{P}(S)$ with minimum distance d and codewords of cardinality at most ℓ is said to be of type $[\log |S|, \log |\mathcal{C}|, d; \ell]$ (we assume that the logarithms are to the base 2, i.e., that the lengths of the messages are measured in bits). The setting we have in mind is the following: The source maps a k -bit information sequence to a set of ℓ n -bit symbols/packets which are sent through the channel. In the channel, these symbols are permuted, some of them are deleted, some of them are received erroneously, and possibly some new symbols are inserted. The receiver collects all these symbols and attempts to reconstruct the information sequence.

Having the above scenario in mind, we can also define the rate of an $[n, k, d; \ell]$ subset code as:

$$R = \frac{k}{n\ell}. \quad (5)$$

A. Isomorphism of subset codes and binary codes

When the ambient set S is specified, the subsets of S are uniquely determined by their characteristic functions (also called indicator functions). The characteristic function of a set $X \subseteq S$ is a mapping $\mathbb{1}_X : S \rightarrow \{0, 1\}$, defined by:

$$\mathbb{1}_X(x) = \begin{cases} 1 & x \in X \\ 0 & x \notin X. \end{cases} \quad (6)$$

If $S = \{1, \dots, q\}$, these functions can be identified with binary sequences of length q , namely $(\mathbb{1}_X(1), \dots, \mathbb{1}_X(q))$. All set operations (unions, intersections, differences, etc.) on $\mathcal{P}(S)$ can be expressed in terms of the corresponding characteristic functions. For example, it is easy to see that:

$$\mathbb{1}_{X \triangle Y} = \mathbb{1}_X \oplus \mathbb{1}_Y = |\mathbb{1}_X - \mathbb{1}_Y| \quad (7)$$

where \oplus denotes the XOR operation (addition modulo 2). Also, the cardinality of the set X can be expressed as:

$$|X| = \sum_x \mathbb{1}_X(x), \quad (8)$$

which is just the Hamming weight of the binary sequence $(\mathbb{1}_X(1), \dots, \mathbb{1}_X(q))$. From the above one concludes that:

$$d(X, Y) = |X \triangle Y| = \sum_x |\mathbb{1}_X(x) - \mathbb{1}_Y(x)|, \quad (9)$$

i.e., the distance between sets X and Y is equal to the Hamming distance between the binary sequences corresponding to $\mathbb{1}_X$ and $\mathbb{1}_Y$. The above reasoning implies the following interesting fact: The subset codes in $\mathcal{P}(S)$ are just a different representation of binary codes in the space $\{0, 1\}^{|S|}$ under the Hamming metric. Every subset code of type $[n, k, d; \ell]$ has a binary counterpart with parameters $(2^n, k, d)$ and maximum codeword weight ℓ , and vice versa.

Example 1: Let $S = \{1, 2, 3, 4, 5\}$. Any subset of S can be identified by a binary sequence of length 5; for example $\{1, 2\} \leftrightarrow 11000$, $\{2, 4\} \leftrightarrow 01010$, etc. Consider now some code in $\{0, 1\}^5$, e.g., $\mathcal{C} = \{11000, 01010, 01110, 00111\}$. The subset counterpart of this code is then $\mathcal{C}_s = \{\{1, 2\}, \{2, 4\}, \{2, 3, 4\}, \{3, 4, 5\}\}$. The distance between two subsets of S is the Hamming distance between the corresponding binary sequences, for example:

$$d(\{1, 2\}, \{2, 4\}) = |\{1, 4\}| = 2 = d_H(11000, 01010) \quad (10)$$

so that all properties of \mathcal{C} directly translate into equivalent properties of the subset code \mathcal{C}_s .

An important consequence of this isomorphism is that subset codes can be constructed by using the familiar constructions of the codes for binary channels. Apart from the construction itself, the analogy can be used for the analysis of the transmission of a subset through a channel. Namely, an equivalent way of describing that X was sent and Y was received, is that the binary word $(\mathbb{1}_X(1), \dots, \mathbb{1}_X(q))$ was sent (through the corresponding binary channel) and $(\mathbb{1}_Y(1), \dots, \mathbb{1}_Y(q))$ was received. Insertion of an element $i \notin X$ to X corresponds to the $0 \rightarrow 1$ transition in the binary channel, i.e., $\mathbb{1}_X(i) = 0$ and $\mathbb{1}_Y(i) = 1$. Similarly, deletion of an element i from X corresponds to the $1 \rightarrow 0$ transition, and a substitution corresponds to both transitions (at different positions) as it is essentially a combination of an insertion and a deletion. Consider further the special case when only deletions can occur in the channel (recall that this is a frequent model of an end-to-end transmission over packet networks). It is easy to conclude from the above discussion that this channel is equivalent to the so-called Z-channel in which the crossover $1 \rightarrow 0$ occurs with probability p (the probability of deletion), while the crossover $0 \rightarrow 1$ never occurs. The analysis of subset codes and the corresponding permutation channel with deletions is thus reduced to the analysis of binary codes and the binary Z-channel, respectively. Note that, for both these channels, we can design a binary code with appropriate parameters. The difference is that, in the binary channel we send a codeword (binary sequence) itself, while

in the subset case, what we send through the channel are the *positions of ones* in this codeword.

IV. MULTISSET CODES

In this section, we generalize subset codes by allowing the codewords to contain multiple copies of their elements. This feature is quite natural, because any interesting classical code over a finite alphabet contains codewords with multiple occurrences of some symbols. In our case, the codewords are sets, and the objects we need – sets with repetitions of elements allowed – are known as *multisets* [6]. A multiset is defined with a set of elements it contains, and numbers of occurrences of each element in the set. The number of occurrences of an element, called its multiplicity, is assumed to be finite. Finally, we note that multisets are also invariant under permutations and hence are suitable for the permutation channel.

Let $\mathcal{M}(S)$ denote the collection of all multisets over an alphabet S . Operations on $\mathcal{M}(S)$, such as union, intersection, difference, etc., are straightforward extensions of the corresponding operations on sets. It is easiest to illustrate them on a simple example.

Example 2: Let $X = \{1, 2, 2, 2, 3\}$ and $Y = \{1, 2, 2, 3, 3, 4\}$ be two multisets over $S = \{1, 2, 3, 4\}$. Then $X \cap Y = \{1, 2, 2, 3\}$, $X \cup Y = \{1, 2, 2, 2, 3, 3, 4\}$, $X \setminus Y = \{2\}$, $Y \setminus X = \{3, 4\}$. The cardinality of X and Y is $|X| = 5$, $|Y| = 6$, respectively.

Codes in the space $\mathcal{M}(S)$ are defined analogously to the codes in $\mathcal{P}(S)$. In the following, $\mathcal{M}(S, \ell)$ denotes the collection of all multisets of cardinality ℓ .

Definition 2: A *multiset code* over S is a nonempty subset of $\mathcal{M}(S)$. If $\mathcal{C} \subseteq \mathcal{M}(S, \ell)$, we say that \mathcal{C} is a constant-cardinality code.

Note that $\mathcal{M}(S)$ is an infinite space. It is always assumed, however, even if not explicitly stated, that a multiset code is finite. In particular, we have in mind multiset codes with an upper bound on the cardinality of the codewords, which is a reasonable constraint from the "practical" point of view. In any case, we shall mostly deal with constant-cardinality codes where this issue does not arise.

It is easy to see that the function d from (1) is a metric on $\mathcal{M}(S)$, and hence we can define the minimum distance of a multiset code in the same way as for subset codes. Other code parameters are also defined in the same way as for subset codes and those definitions will not be repeated here.

We now prove a simple, but basic fact about the correcting capabilities of multiset codes. The analogous statement for the special case of subset codes is proven in [1].

Theorem 3: A multiset code \mathcal{C} with minimum distance d is capable of correcting any pattern of s insertions, ρ deletions, and t substitutions, as long as $2(s + \rho + 2t) < d$.

Proof: Let $X \in \mathcal{C}$ be the multiset which is transmitted through the channel. Let Y be the received multiset. If ρ packets from X have been deleted, and s new packets have been inserted, then we easily deduce that $|X \cap Y| \geq |X| - \rho$ and $|Y| = |X| - \rho + s$. Since each substitution is essentially

a combination of one deletion and one insertion, the actual number of deletions and insertions is $\rho + t$ and $s + t$, respectively, wherefrom one concludes that $|X \cap Y| \geq |X| - \rho - t$ and $|Y| = |X| - \rho + s$, and that

$$d(X, Y) = |X| + |Y| - 2|X \cap Y| \leq s + \rho + 2t. \quad (11)$$

Now, if the assumption $2(s + \rho + 2t) < d$ holds, then $d(X, Y) \leq \lfloor \frac{d-1}{2} \rfloor$ and therefore X can be recovered from Y by the minimum distance decoder. ■

If only deletions can occur in the channel, then $d(X, Y) = \rho$ and the sent codeword is recoverable whenever $\rho \leq \lfloor \frac{d-1}{2} \rfloor$.

An obvious advantage that multiset codes have over subset codes is the code rate improvement which is a consequence of them being defined in a bigger space:

$$|\mathcal{M}(S, \ell)| = \binom{q + \ell - 1}{\ell} > \binom{q}{\ell} = |\mathcal{P}(S, \ell)|. \quad (12)$$

Further, when $|S| = q$ is "small", it is necessary to use multiset codes because, unlike subset codes, they allow the cardinality of the codewords to be larger than the cardinality of the alphabet. For example, multiset codes with arbitrary minimum distance (and hence, arbitrary correction capability) can be defined even over a binary alphabet.

A. Isomorphism of multiset codes and integer codes

The isomorphism between subset codes and binary codes, which has many important consequences, as discussed in Section III-A, also has an appropriate generalization in the multiset framework. Namely, multiset codes turn out to be equivalent to integer codes under the so-called Manhattan metric, and this equivalence is illustrated next.

Multisets over an alphabet S can be described by their *multiplicity functions* in the same way subsets are described by their characteristic functions (in fact, that is how multisets are usually defined formally [6]). The multiplicity function of a multiset X over S is a mapping $\mathfrak{m}_X : S \rightarrow \mathbb{Z}_{\geq 0}$, where $\mathfrak{m}_X(x)$ is the number of occurrences of x in X . Clearly, a multiset is a set if and only if the range of its multiplicity function is $\{0, 1\}$. Operations on multisets can be expressed in terms of their multiplicity functions, for example:

$$\begin{aligned} \mathfrak{m}_{X \cup Y} &= \max\{\mathfrak{m}_X, \mathfrak{m}_Y\}, \\ \mathfrak{m}_{X \cap Y} &= \min\{\mathfrak{m}_X, \mathfrak{m}_Y\}, \\ \mathfrak{m}_{X \setminus Y} &= \max\{0, \mathfrak{m}_X - \mathfrak{m}_Y\}, \end{aligned} \quad (13)$$

while the cardinality of a multiset is expressed as:

$$|X| = \sum_x \mathfrak{m}_X(x). \quad (14)$$

If the alphabet is $S = \{1, 2, \dots, q\}$, the multiplicity function of a multiset X is uniquely specified by a sequence $(\mathfrak{m}_X(1), \dots, \mathfrak{m}_X(q)) \in \mathbb{Z}_{\geq 0}^q$. Therefore, the space $\mathcal{M}(S)$ is essentially equivalent to the space $\mathbb{Z}_{\geq 0}^q$. Further, the distance between multisets is:

$$d(X, Y) = |X \triangle Y| = \sum_x |\mathfrak{m}_X(x) - \mathfrak{m}_Y(x)|, \quad (15)$$

which is the familiar ℓ_1 distance, also known as the Manhattan metric. Therefore, multiset codes are basically just another

description of the codes in $\mathbb{Z}_{\geq 0}^q$ under the Manhattan metric. Constant-cardinality codes are then equivalent to the codes on the "sphere" $\{(x_1, \dots, x_q) : x_i \in \mathbb{Z}_{\geq 0}, \sum_i x_i = \ell\}$.

V. EXAMPLES OF CODES FOR THE PERMUTATION CHANNEL

In this section, we describe a simple way to construct subset and multiset codes, and discuss some of the properties of the obtained codes.

A. Example of subset codes

A straightforward way of obtaining codes for the permutation channel is to use some classical error-correcting code, and add a sequence number to every symbol of the codeword so that the order of symbols can be restored at the receiving side. This approach is illustrated below.

Let \mathcal{A} be a finite alphabet with $|\mathcal{A}| = q$. Observe some code \mathcal{C} over \mathcal{A} with parameters (ℓ, k, d) , meaning that $|\mathcal{C}| = q^k$, the codewords of \mathcal{C} are q -ary sequences of length ℓ , and the Hamming distance between any two codewords is at least d . For any codeword $\mathbf{p} = (p_1, \dots, p_\ell) \in \mathcal{C}$, create a sequence (t_1, \dots, t_ℓ) , where $t_i = i \circ p_i$ is a new symbol obtained by prepending a sequence number to the symbol p_i (\circ denotes the concatenation of strings). This mapping is clearly injective and the set of all sequences thus obtained defines a code \mathcal{C}' over an alphabet $S = \{1, \dots, \ell\} \times \mathcal{A}$ with parameters (ℓ, k, d) . The codewords of \mathcal{C}' are invariant under permutations, i.e., any permutation of (t_1, \dots, t_ℓ) has the same meaning to the receiver because it can recover (p_1, \dots, p_ℓ) from the sequence numbers. Therefore, one can imagine the carrier of information being a *set* $\{t_1, \dots, t_\ell\}$, and hence this simple construction yields an example of a subset code \mathcal{C}_s over S . The code has q^k codewords, each of cardinality ℓ . The minimum (subset) distance of the code is easily determined by observing two codewords:

$$\begin{array}{cccc} 1 \circ p_1 & 2 \circ p_2 & \dots & \ell \circ p_\ell \\ 1 \circ r_1 & 2 \circ r_2 & \dots & \ell \circ r_\ell. \end{array} \quad (16)$$

It is evident that the cardinality of the intersection of the subset codewords $P = \{1 \circ p_1, \dots, \ell \circ p_\ell\}$ and $R = \{1 \circ r_1, \dots, \ell \circ r_\ell\}$ is equal to the number of positions where the sequences $\mathbf{p} = (p_1, \dots, p_\ell)$ and $\mathbf{r} = (r_1, \dots, r_\ell)$ agree, which is, on the other hand, equal to ℓ minus the Hamming distance of these two sequences. Therefore,

$$d(P, R) = 2d_H(\mathbf{p}, \mathbf{r}), \quad (17)$$

and hence the minimum (subset) distance of \mathcal{C}_s is $2d$. To conclude, this construction yields a subset code \mathcal{C}_s of type $[\log q \ell, k \log q, 2d; \ell]$.

Note that the decoding procedure for \mathcal{C}_s is the same as for \mathcal{C} once the codeword of \mathcal{C} is recovered by using sequence numbers. Note also that recovering (p_1, \dots, p_ℓ) from $\{1 \circ p_1, \dots, \ell \circ p_\ell\}$ reduces deletions to *erasures*, while insertions and substitutions are reduced to *errors*. Namely, if $i \circ p_i$ has been deleted, the receiver will be able to deduce that the symbol at the i th position is missing. Similarly, if $j \circ p_j$ has been inserted and the receiver now possesses two symbols

with the sequence number j , it will choose one at random, possibly resulting in an error at the j th position. Hence, when subset codes constructed in this way are used, the permutation channel (over S) with insertions, deletions, and substitutions, reduces to the classical discrete memoryless channel (over \mathcal{A}) with errors and erasures.

The codes described above are, to the best of our knowledge, the only type of error-correcting codes for the permutation channel described in the literature (see, e.g., the construction of the "outer" code in [4]). As we have illustrated, they are in fact only a special case of the more general notion of subset codes. We note that better subset codes can be constructed via the isomorphism given in Section III-A, i.e., by using the familiar constructions of binary codes (see also [1]).

B. Example of multiset codes

We next describe a simple construction which yields an example of a multiset code (which is not a subset code). It is also based on "classical" codes and sequence numbers.

Again, let \mathcal{A} be a finite alphabet with $|\mathcal{A}| = q$ symbols, and \mathcal{C} a code over \mathcal{A} . For any codeword $\mathbf{p} = (p_1, \dots, p_\ell) \in \mathcal{C}$, we create a sequence (t_1, \dots, t_ℓ) by prepending sequence numbers to the symbols of \mathbf{p} , but in such a way that runs of identical symbols in \mathbf{p} are given the same sequence number. For example, the sequence (a, a, b, b, c, b) , where $a, b, c \in \mathcal{A}$, is mapped to $(1 \circ a, 1 \circ a, 2 \circ b, 2 \circ b, 3 \circ c, 4 \circ b)$. The obtained sequence is invariant under permutations, and it is easily concluded, similarly to the example from the previous subsection, that this procedure yields a multiset code \mathcal{C}_M over S . The decoding procedure for \mathcal{C}_M is again the same as for \mathcal{C} once the codeword is recovered from the sequence numbers. In this case, however, recovering \mathbf{p} from $\{i_1 \circ p_1, \dots, i_\ell \circ p_\ell\}$ reduces deletions to *deletions*, insertions to either *insertions* or *substitutions*, and substitutions to *substitutions* (i.e., *errors*). Namely, if the symbol $i_j \circ p_j$ has been deleted, the receiver cannot deduce (in general) which symbol has been deleted because there could have been multiple copies of this or some other symbols. Similar reasoning applies for the other cases. Therefore, the code \mathcal{C} has to be resilient to insertions, deletions, and substitutions.

Finally, let us determine the parameters of \mathcal{C}_M from those of \mathcal{C} . Let \mathcal{C} be of type (ℓ, k, d) , where k and ℓ are as before, and d is the minimum Levenshtein distance [7], which is the relevant distance measure for insertion/deletion channels (it is defined as the minimum number of insertions and deletions that transform one sequence to the other). Observe two multiset codewords P and R :

$$\begin{array}{ccccccc} i_1 \circ p_1 & i_2 \circ p_2 & \dots & i_\ell \circ p_\ell & & & \\ j_1 \circ r_1 & j_2 \circ r_2 & \dots & j_\ell \circ r_\ell, & & & \end{array} \quad (18)$$

where, (i_m) and (j_m) are nondecreasing integer sequences, as explained above. Unfortunately, the distance between P and R in general cannot be expressed via the Levenshtein (or Hamming) distance between \mathbf{p} and \mathbf{r} , and hence the minimum distance of \mathcal{C}_M cannot be inferred from d . It is easy to conclude, however, that the distance between two multisets obtained in

this way is greater than or equal to the Levenshtein distance between the original sequences, and therefore the code \mathcal{C}_M is of type $[\log q\ell, k \log q, d_M; \ell]$, where $d_M \geq d$.

As noted above, one possible decoding procedure for \mathcal{C}_M is to first use the sequence numbers to obtain the right ordering of symbols, and then apply the decoding algorithm for \mathcal{C} to the resulting sequence. If this procedure is used, then one easily concludes that the number of insertions and deletions which can be corrected is at most $\lfloor \frac{d-1}{2} \rfloor$, and therefore, the "effective minimum distance" of the code is d .

As a final note here, we would like to stress that the above construction merely serves as an illustration of a constant-cardinality multiset code. The general method of construction that can be used is via the corresponding constant-weight integer codes in the Manhattan metric, as explained in Section IV-A. It appears, however, that these codes have not been studied thoroughly before, and it remains an interesting problem for future research to explore further their properties, and obtain explicit constructions and decoding algorithms.

VI. CONCLUSION

We have presented a framework for forward error correction in the permutation channels. We have introduced multiset codes as relevant constructs for correcting insertions, deletions, and substitutions in such channels. Some basic properties of multiset codes have been established. The framework presented is analogous to the one introduced recently by Kötter and Kschischang for the operator channels, and can be viewed as its extension. As a consequence, a unified view on coding for RLNC networks and multipath routed packet networks is obtained.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of the Ministry of Science and Technological Development of the Republic of Serbia (grants No. TR32040 and III44003).

REFERENCES

- [1] M. Kovačević and D. Vukobratović, "Subset Codes for Packet Networks," *IEEE Commun. Lett.*, submitted for publication, revised Jan. 2013, available at arXiv:1210.7341.
- [2] D. P. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Prentice Hall, 1992.
- [3] R. Kötter and F. R. Kschischang, "Coding for Errors and Erasures in Random Network Coding," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3579–3591, Aug. 2008.
- [4] L. J. Schulman and D. Zuckerman, "Asymptotically Good Codes Correcting Insertions, Deletions, and Transpositions," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2552–2557, Nov. 1999.
- [5] G. Birkhoff, *Lattice Theory*, 3rd ed., American Mathematical Society, 1967.
- [6] M. Aigner, *Combinatorial Theory*, Springer, 1979.
- [7] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Sov. Phys.-Dokl.*, vol. 10, no. 8, pp. 707–710, Feb. 1966.